Evaluating ML models for lightning forecasting in Brazil.

Arielle dos Santos Bassanelli Pereira[1], Álvaro L. Fazenda[1], Alan James Peixoto Calheiros[2]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (Unifesp)
São José dos Campos – SP

[2]Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP

{arielle.bassanelli,alvaro.fazenda}@unifesp.br, alan.calheiros@inpe.br

Abstract. *Instruments for monitoring severe meteorological phenomena (such as lightning, flooding and landslides) can be used to assist in decision-making by state agencies, in an attempt to mitigate their possible harmful effects. These phenomena usually occur suddenly on a short-term duration, under a limited region, imposing difficulties in being predicted by regular weather forecast models, requiring specific prediction systems. Very short-term weather forecasting systems, on order of a few hours, known as nowcasting, can include numerical models of physical phenomena and machine learning algorithms. This work presents a system for forecasting the incidence of lightning, a common phenomenon in electrically active storms, through the application and evaluation of two machine learning models, an Artificial Neural Network and a Random Forest model, which were able to detect the occurrence of atmospheric electrical discharges from the automatic recognition of patterns obtained from the data generated by the numerical weather forecasts. The Random Forest model presented the best results when trained with the set that includes the ten best correlated variables, reaching 99.77% of accuracy for the case study performed.*

1. Introduction

To carry out a short-term weather forecast (known as nowcasting), the operational weather and climate forecasting centers use past and real-time information about the precipitation intensity, electrical discharges, and cloudiness occurring in the monitored locations [Sakuragi 2017]. This meteorological data can be obtained through radars, satellites, automatic weather stations, or even through data produced through a - Numerical Weather Prediction (NWP) model. The alerts' reliability from the operational monitoring centers depends on the spatial resolution and updating frequency from the meteorological observations available in an integrated display system, which can be easily operated by the forecaster, in addition to the forecaster's knowledge and experience [WMO 2017].

This work aims to assist the forecaster in monitoring and predicting severe weather events, automatically identifying regions associated with the development of storms that produce electrical discharges (lightning), through a Machine Learning algorithms (ML) [Faceli et al. 2011]. Machine learning is an sub-area of Artificial Intelligence (AI) related to the development of software that explore techniques capable of finding their own solution, learning through examples, that is, simulating characteristics present in human rational behavior, relying on the advantages of time, robustness and reliability that a computational system can offer, and using inductive reasoning to conclude a result after considering a sufficient number of particular cases. In the last decade Nowcasting has increasingly benefited from rapid advances in ML models [Leinonen et al. 2022].

Two methods were chosen to recognize patterns that characterize atmospheric conditions: Artificial Neural Network (ANN) with the the well-known Multi-Layer Perceptron (MLP) [Haykin 2001] and Random Forest (RF) [Ho 1995]. The training phase for both used estimated electrical discharge data combined with forecast data from an NWP model, joined with an algorithm to extract the data developed during the research.

2. Materials, Methods and Methodology

As input data for training and testing ML methods to classify and predict the intensity of electrical activity in the atmosphere, data from the Weather Research and Forecasting Model (WRF) [NCAR 1990] were used. Also, data from Geostationary Lightning Mapper (GLM) sensor have been used to define the atmospheric events severity. The GLM sensors are located on GOES-16 satellite in a geostationary orbit, and represent a single-channel, near-infrared, optical transient detector. It also detects continuously all forms of lightning on day and night with 10km for spatial resolution and 5Km for mapping accuracy, with detection efficiency of 70% or better with a False Alarm Rate (FAR) $<$ 5% [NOAA/NESDIS 2012].

The WRF raw data were provided in NetCDF [Rew and Davis 1990] format, with approximately 110 GB for each simulation, totaling 4.9 TB of data. For the observed data, data from the GOES 16 satellite were used, with the GLM sensor, pre-processed by INPE (National Institute for Space Research in Brazil) for the period between 01/02/2019 at 00:00 UTC and 01/24/2019 at 23:00 UTC, providing a new data grid every 15 minutes for South America, with a spatial resolution of 8 x 8 km. This data was also provided in the NetCDF [Rew and Davis 1990] format, with approximately 90 Kb per file, totaling 65 GB.

The Region of Interest (RoI) was defined over South America. Ideally, in nowcasting systems the forecast models are started for every hour, however, the operational WRF model used was started only on 12-hour intervals, with a 1-hour output frequency. So for the interval between 0 am and 11 am the output data generated from the 00Z model start was used. For the remaining interval, from 0 pm to 11 pm, the data generated by the model started on 12Z have been used. The GLM data was extracted by looking for samples with positive flash value (the flash information is based on the density of lights per km²). Therefore, for the same instant, tuples (WRF,GLM) were generated, with and without lightning. These data for each time step were the basis for the ML's training dataset.

Two types of classification were defined for the intensity of lightning:

- Binary Classification - dataset is divided into two groups: inactivity (class 0) and flash activity (class 1);
- Multi-class - data referring to the class of inactivity in the occurrence of lightning (class 0), and data grouped into three groups (clusters), defining different classes of lightning from GLM, representing low (class 1), medium (class 2) and high (class 3) flash activity, with their thresholds defined using the K-Means [MacQueen 1967] method with three clusters.

To identify which atmospheric output variables from the WRF model shows the major influence on the occurrence of lightning, a correlation was calculated between all the WRF output variables and the lightning density/km² from the GLM. From them, four sets with different numbers of parameters/variables to be used for ML training were defined:

- Dataset 1 - The first 5 variables with the highest correlation value;
- Dataset 2 - The first 10 variables with the highest correlation value;
- Dataset 3 - The first 15 variables with the highest correlation value;
- Dataset 4 - A set of 12 variables with the highest correlation values chosen from empirical knowledge of meteorology and computer scientists. The idea is to verify whether the data chosen based on knowledge about meteorology may suffer from the choices of non-meteorologists.

In this research, three ML models were used: MLP-type ANN trained for binary classification, labeled as regions with and without lightning; MLP-type ANN trained to classify the severity of these phenomena (no flash activity, little activity, medium activity and high activity); and RF, also trained to classify the severity of the events. The ML models were implemented using Python's *sckit-learn* [Pedregosa et al. 2011] framework.

3. Results

In the data extraction over the RoI for the period, a total of 15,372 flash samples were detected and selected, joined with the same number of inactivity points in different regions to keep the data used in the ML methods balanced, totaling 30,743 samples.

Almost all variables representing wind, temperature and humidity at various vertical levels was included in the fifteen best WRF variables correlated to GLM data, showing a recurring pattern. According to a Systematic Literature Review performed by the authors in the field of interest, it was noted that the four variables most used to predict meteorological events in studies with AI are precipitation, wind, temperature, and humidity. Therefore the correlation presented a result including variants of the meteorological information used in 60% of studies according to the same review.

To analyze the performance the following metrics was used:

1. POD (Probability of Detection):

$$POD = \frac{TruePositives}{(TruePositives + FalsiesNegatives)} \tag{1}$$

2. Precision:

$$precision = \frac{TruePositives}{(TruePositives + FalsiesPositives)} \tag{2}$$

3. FAR (false-alarm rate):

$$FAR = \frac{FalsiesPositives}{(TruePositives + FalsiesNegatives)} \tag{3}$$

4. Bias [Luxburg and Scholkopf 2011]:

$$BIAS = \frac{(TruePositives + FalsiesPositives)}{(TruePositives + FalsiesNegatives)} \tag{4}$$

Based on the Table 1, all models showed satisfactory results, except for Binary ANN and MC ANN with dataset 4. Also, in Table 1 it is observed that set 2 presented the best performance among the ML models studied, which contain the ten WRF variables best correlated with GLM data. For the evaluated period, most of the models trained with set 2 successfully achieved the lightning activity and inactivity pattern identification, hitting 100% of the cases. It should be noted that the results of the validations are associated with a sample set of data and not with the total points of the WRF model's output grid, which was a limitation of the study strategy.

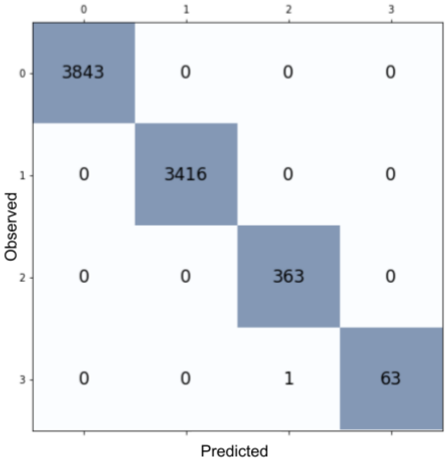
Tabela 1. ML Models Results for the 4 sets.

ML Model	Set	FAR	Precision	POD	Bias
Binary ANN	1	0,0000	1,0000	1,0000	1,0000
Binary ANN	2	0,0000	1,0000	1,0000	1,0000
Binary ANN	3	0,0010	0,9990	1,0000	1,0010
Binary ANN	4	0,7827	0,2173	0,6850	0,3153
MC ANN	1	0,0003	0,9997	1,0000	1,0003
MC ANN	2	0,0000	1,0000	1,0000	1,0000
MC ANN	3	0,0003	0,9997	0,9997	1,0000
MC ANN	4	0,0000	1,0000	0,5296	0,5296
RF	1	0,0000	1,0000	1,0000	1,0000
RF	2	0,0000	1,0000	1,0000	1,0000
RF	3	0,0000	1,0000	1,0000	1,0000
RF	4	0,0000	1,0000	1,0000	1,0000

The hypothesis that set 2 is more effective is due to the inclusion of a humidity profile up to average levels in the atmosphere, which is an essential ingredient for cloud formation in pre-convective situations [Cotton et al. 2010]. Although set 3 also presented the same profile, the selection of more variables should be less effective, introducing more errors in the forecast. Set 1 was limited to regions closer to the surface, which did not represent the layers where the separation of electrical charges is more evident [Rakov and Uman 2003]. And last but not least, the empirical variables choice, which presented the worst result, should be a consequence of good and bad choices, which shows that even using good variables, the selection of less related variables negatively impacts the results.

Figure 1 shows the confusion matrix referring to the lightning forecast for the four defined classes in a specif day (09/01/2019), when high atmospheric electric activity was noted. In this figure, it is possible to perceive the accuracy of the prediction through the RF model using set 2, where the diagonal highlighted cells in gray represents the assertiveness lightning severity class prediction, demonstrating the quantities of predicted

Figura 1. Confusion Matrix for RF model with Set 2 in 09/01/2019.



	0	1	2	3
0	3843	0	0	0
1	0	3416	0	0
2	0	0	363	0
3	0	0	1	63

Observed
Predicted

and observed samples for the classes. Thus, it is possible to notice 3843 samples without atmospheric electrical activity (class 0), 3416 samples with low activity (class 1), 363 samples with moderate activity (class 2), and 63 samples with high activity (class 3) were predicted and observed. It is also important to note the erroneous prediction for only a single sample in the last row of the matrix, where the ML method had predicted high activity instead of an observed moderate activity.

4. Conclusion

In this study, three AM models have been analyzed: ANN with binary classification, ANN multi-classes, and Random Forest multi-classes, for four different training dataset. The four most used variables to predict severe weather events in AI studies, according to a briefly bibliographic review, are precipitation, wind, temperature, and humidity. From the methodology developed for data extraction and the calculation of the correlation between WRF output variables and the GLM data, the best-correlated variables that make up the sets used in the training of the models are concentrated in wind speed variants, temperature, and humidity at various vertical levels.

It is important to note that the presented validation results are strongly associated with a specific period, using a sample set geographically distributed in the RoI, and do not cover the total grid points in the horizontal grid used in the WRF model. Thus, there is a chance that the high degree of assertiveness found is related to specific atmospheric conditions in the period and strongly correlated with the occurrence of electrical activity, which is a limitation of the method used.

It was also observed that ML models presented satisfactory results for short-term prediction. Since they are probabilistic models, the physical and dynamic conditions of the atmosphere over time was not considered. The use of better correlated variables for the construction of data sets proved to be effective.

The data used in this work represent only one month of information in a specific year. This fact suggests that the results of the models created can only represent the events that occurred in this period. In order to generalize the results, it is necessary to train the algorithms for longer periods taking into account the seasonality of convective events in different regional regimes. Such a study would demand greater computational resources and more detailed analyses, which is intended to be carried out in the future.

Acknowledgement

CNPq process 438310/2018-7.

INCT of the Future Internet for Smart Cities funded by CNPq #465446/2014-0, CAPES – Finance Code #001, FAPESP #14/50937-1 and #15/24485-9.

Referências

Cotton, W. R., Bryan, G. H., and C., V. d. H. S. (2010). *Storm and cloud dynamics*. Elsevier, 2 edition.

Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. d. L. F. d. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.

Haykin, S. (2001). *Redes Neurais - Príncipios e Práticas*. Bookman.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*.

Leinonen, J., Hamann, U., Germann, U., and Mecikalski, J. R. (2022). Nowcasting thunderstorm hazards using machine learning: the impact of data sources on performance. *Natural Hazards and Earth System Sciences*, 22(2):577–597.

Luxburg, U. v. and Scholkopf, B. (2011). Statistical learning theory: Models, concepts, and results. *Handbook of the History of Logic*, page 651–706.

MacQueen, J. B. (1967). *Some methods for classification and analysis of Multivariate Observations*, volume 1. Defense Technical Information Center.

NCAR (1990). Weather research and forecasting model. `https://www.mmm.ucar.edu/weather-research-and-forecasting-model`. Accessed: 25/05/2020.

NOAA/NESDIS (2012). Glm lightning cluster-filter algorithm. `https://www.star.nesdis.noaa.gov/goesr/documents/ATBDs/Baseline/ATBD_GOES-R_GLM_v3.0_Jul2012.pdf`. Accessed: 17/06/2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and et al (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rakov, V. A. and Uman, M. A. (2003). Lightning: physics and effects. *Cambridge University*, page 687.

Rew, R. K. and Davis, G. P. (1990). *NetCDF: An Interface for Scientific Data Access*, 10(4):76–82.

Sakuragi, J. (2017). *Estudo da morfologia das tempestades severas em 3D e potencial aplicação em Nowcasting*. PhD thesis, Instituto Nacional de Pesquisas Espaciais – INPE.

WMO (2017). *Guidelines for Nowcasting Techniques*. World Meteorological Organization.